Homework 1

Welcome to the first homework assignment!

The purpose of this homework is to practice using R and R Markdown, and to review some concepts from introductory Statistics. Please fill in the appropriate R and R Markdown and write answers to all questions in the answer section., then submit a compiled pdf or html with your answers to Canvas by 11:59pm on Sunday September 8th.

If you need help with any of the homework assignments, please attend the TA office hours which are listed on Canvas and/or ask questions on Piazza. Also, if you have completed the homework, please help others out by answering questions on Piazza.

Problem 1: RMarkdown practice

RMarkdown has a number of features that allow the text in your written reports to have better formatting. In the following exercise, please modify lines of text to change their formatting. A cheatsheet for RMarkdown formatting can be found here. When answering the questions (i.e., formatting the text below) be sure to knit your RMarkdown document very often to catch errors as soon as they are made.

Problem 1.1: Please format the lines of text below (15 points)

Make this line bold

Make this line italics

Make this line a third level header

- Make this line a bullet point

LINK (make the word LINK at left link to Yale's home page)

Problem 1.2: Use LaTeX to write plato's name in Greek below (10 points)

Note: make sure the ending dollar sign touches the last letter otherwise you will get an error when knitting.

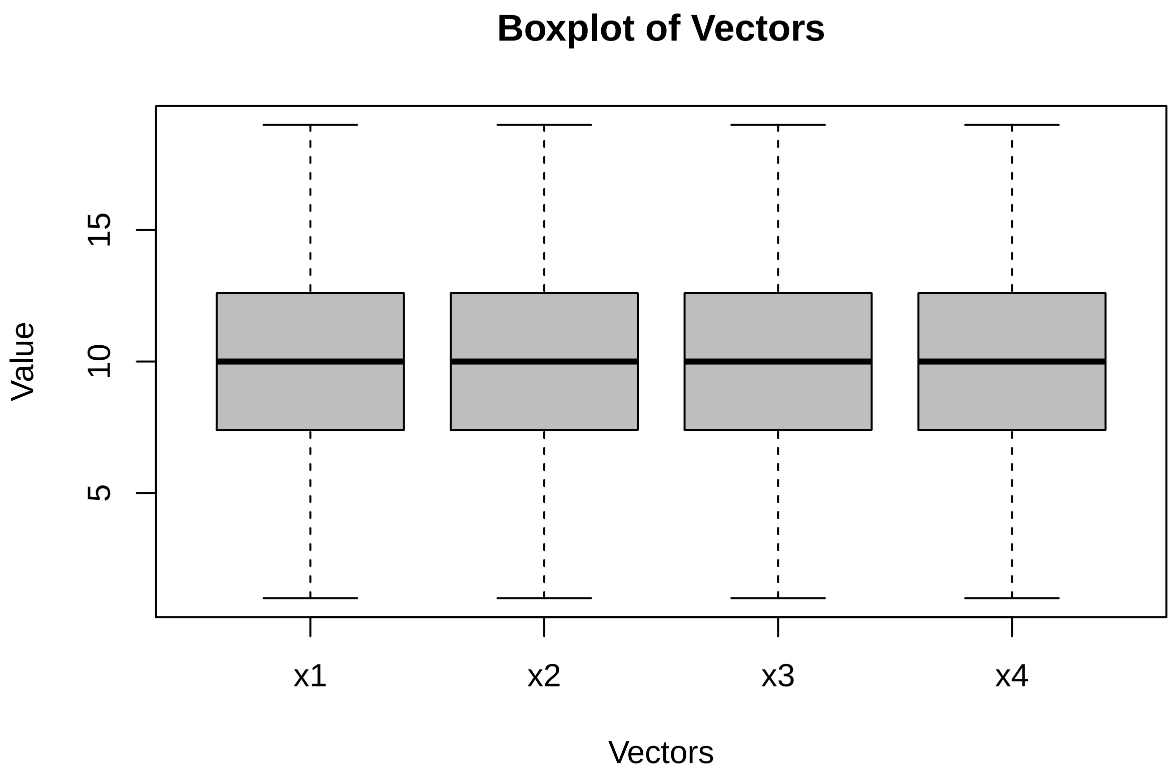
$\Pi\lambda\acute{\alpha}\tau\omega\upsilon$

Problem 2: Descriptive statistics and plots

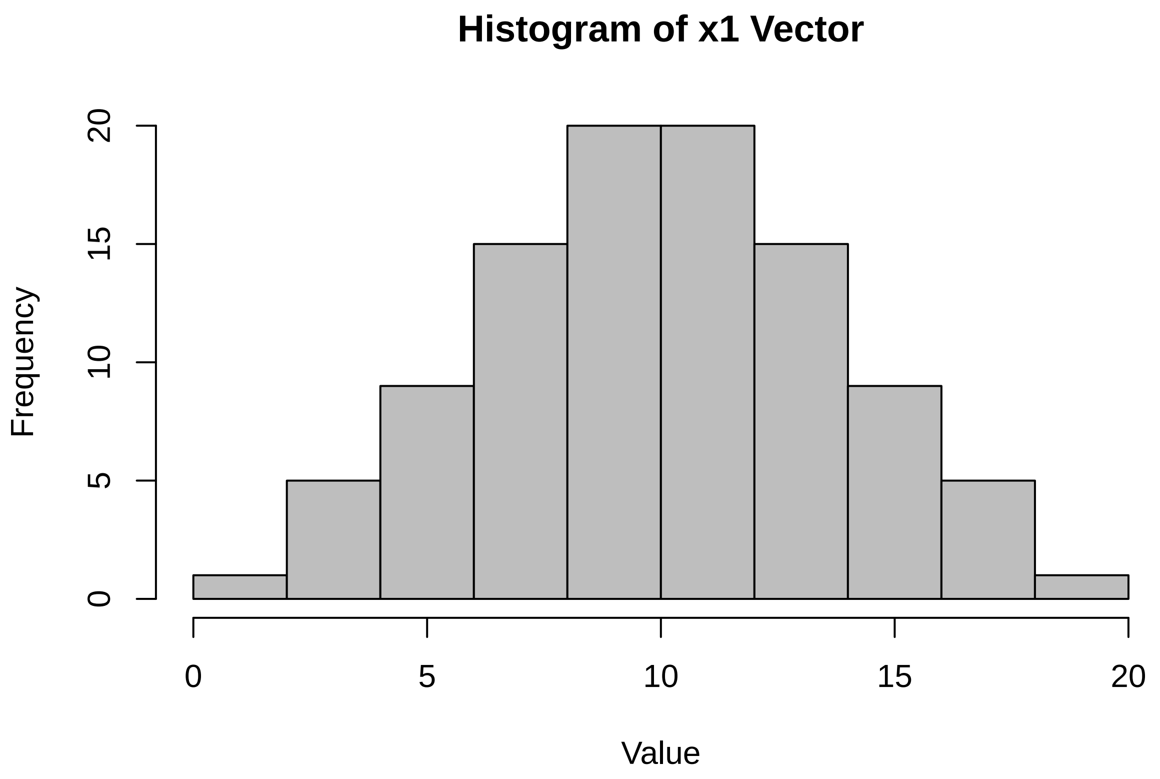
Below you will create and compare a few plots. Please answer each question, and if you notice any outliers in your data please address them appropriately. Also be sure to label your plots appropriately.

Part 2.1: (10 points) The code chunk below loads four vectors objects named x1, x2, x3, and x4. Create a side-by-side boxplot that compares these four vectors. Also create a histogram for each of these vectors (4 histograms total). Describe below whether the boxplots or histograms are more informative for plotting this data and why.

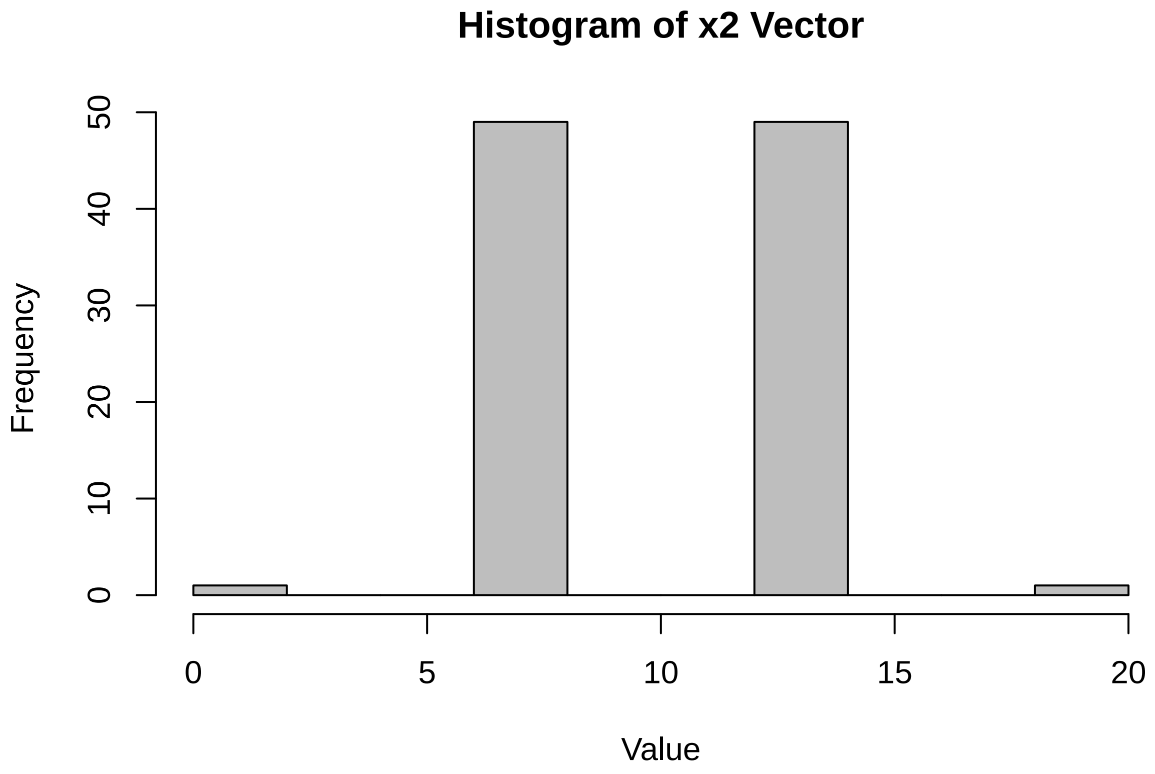
```r
load("misc_data.Rda")
boxplot(x1, x2, x3, x4, names = c("x1","x2", "x3", "x4"), xlab = "Vectors",
        ylab = "Value", main = "Boxplot of Vectors", col = "gray")
```



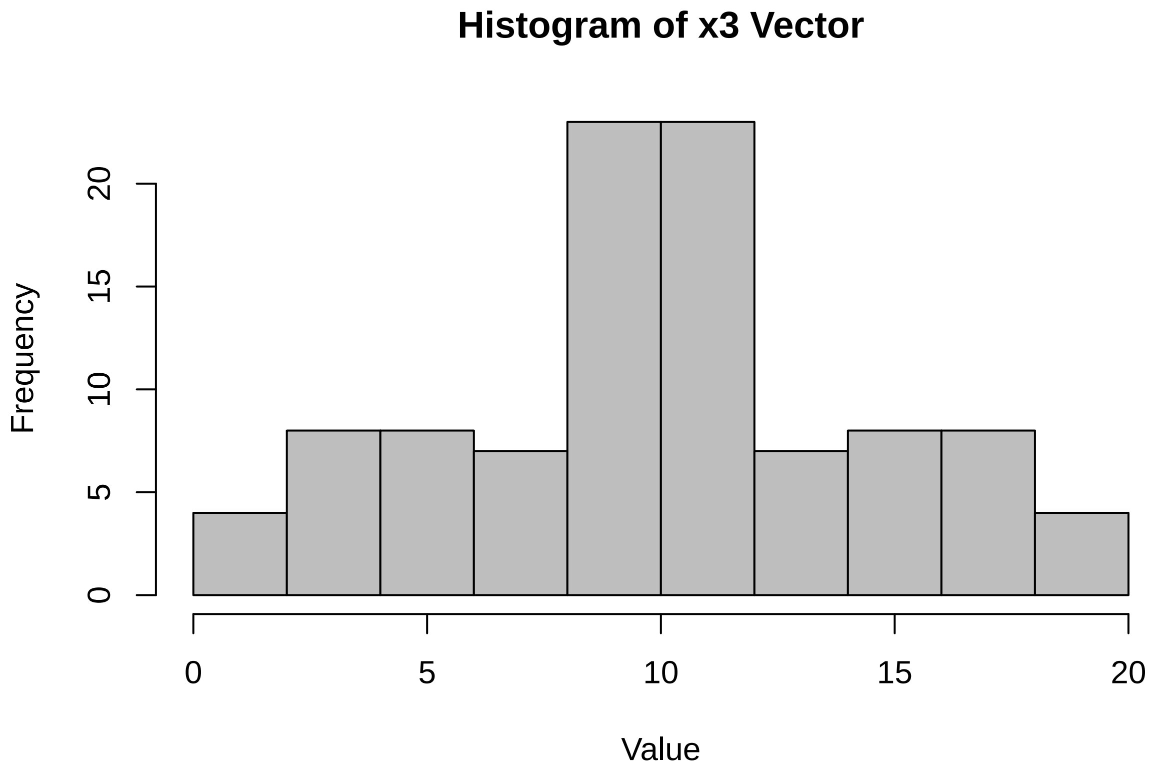
```
hist(x1, xlab = "Value", ylab = "Frequency", main = "Histogram of x1 Vector", col = "gray")
```



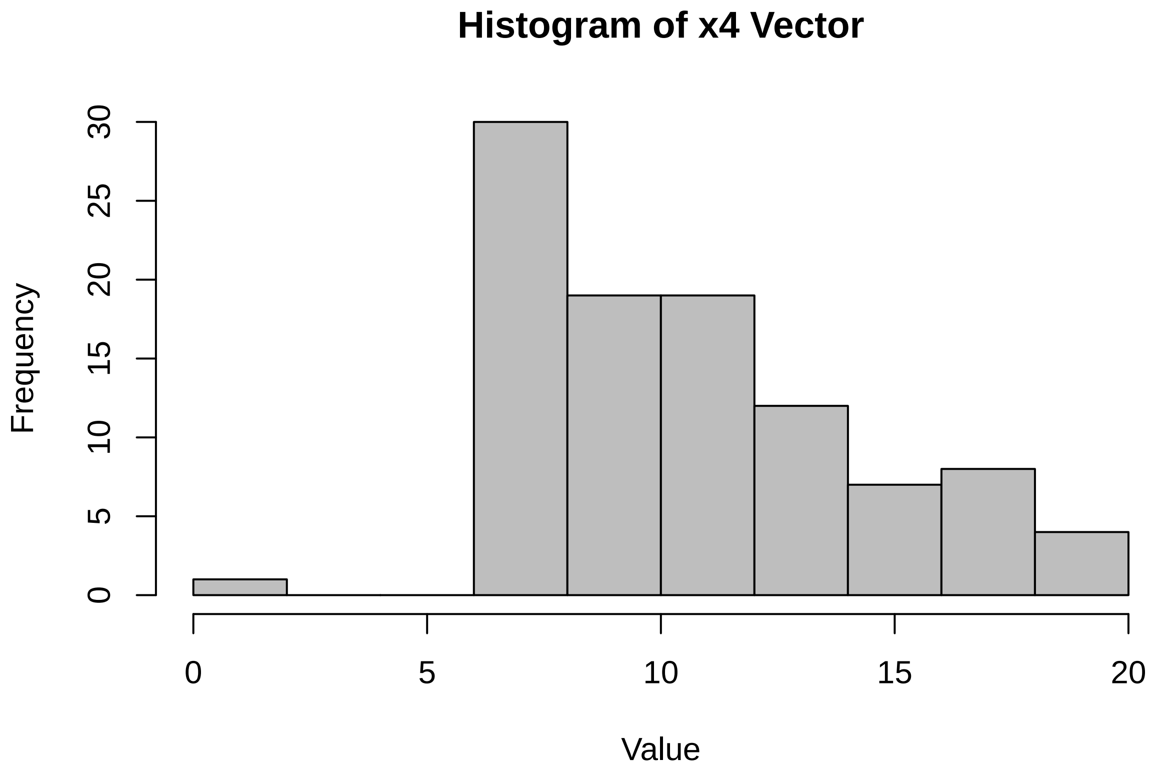
```
hist(x2, xlab = "Value", ylab = "Frequency", main = "Histogram of x2 Vector", col = "gray")
```



```r
hist(x3, xlab = "Value", ylab = "Frequency", main = "Histogram of x3 Vector", col = "gray")
```



```r
hist(x4, xlab = "Value", ylab = "Frequency", main = "Histogram of x4 Vector", col = "gray")
```



Answer: [Describe whether boxplots or histograms are more informative here]

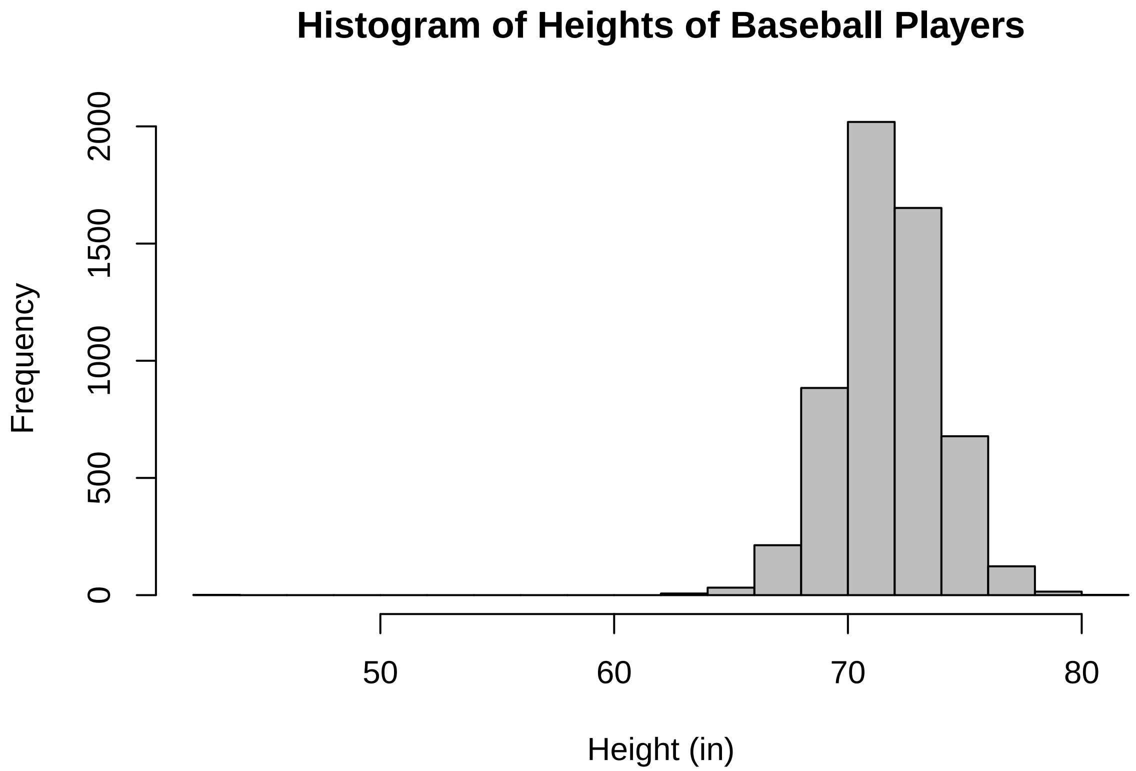
The histograms are more informative here. With the boxplot, we can only see summary statistics such as the median, quartiles, and extremes. In regards to this data set, for x1, x2, x3, and x4, all of these values are exactly the same. However, with the histograms, we can see the actual distribution of data points. The histograms reveal that the data sets are actually quite different despite having the same summary statistics.

Part 2.2: (10 points) The R chunk below loads a data frame with information on all major league baseball players who were born between 1901 and 1950 (if you are interested in the data, it comes from the Lahman package). Create a vector object that is called `heights`, that has just the player heights. Then create a histogram and a boxplot of the players' heights using this vector object. Describe the shape of the distribution of heights and any advantages that one type of plot has over the other. Also investigate any unusual features of the data.

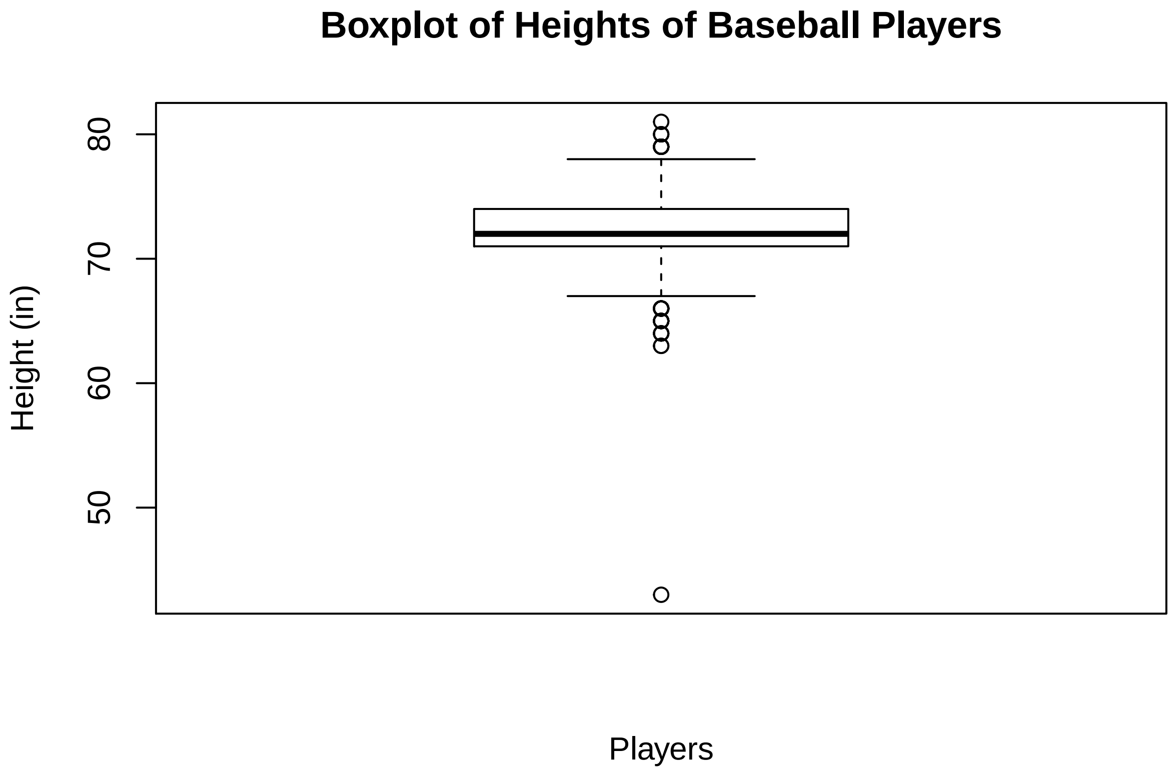
```
load("players_born_1901_1950.Rda")
heights <- players_born_1901_1950$height
mean(heights, na.rm = TRUE)
```

```
## [1] 72.23307
```

```
hist(heights, main = "Histogram of Heights of Baseball Players", xlab = "Height (in)",
      ylab = "Frequency", col = "gray")
```



```
boxplot(heights, main = "Boxplot of Heights of Baseball Players", ylab = "Height (in)",
      xlab = "Players")
```

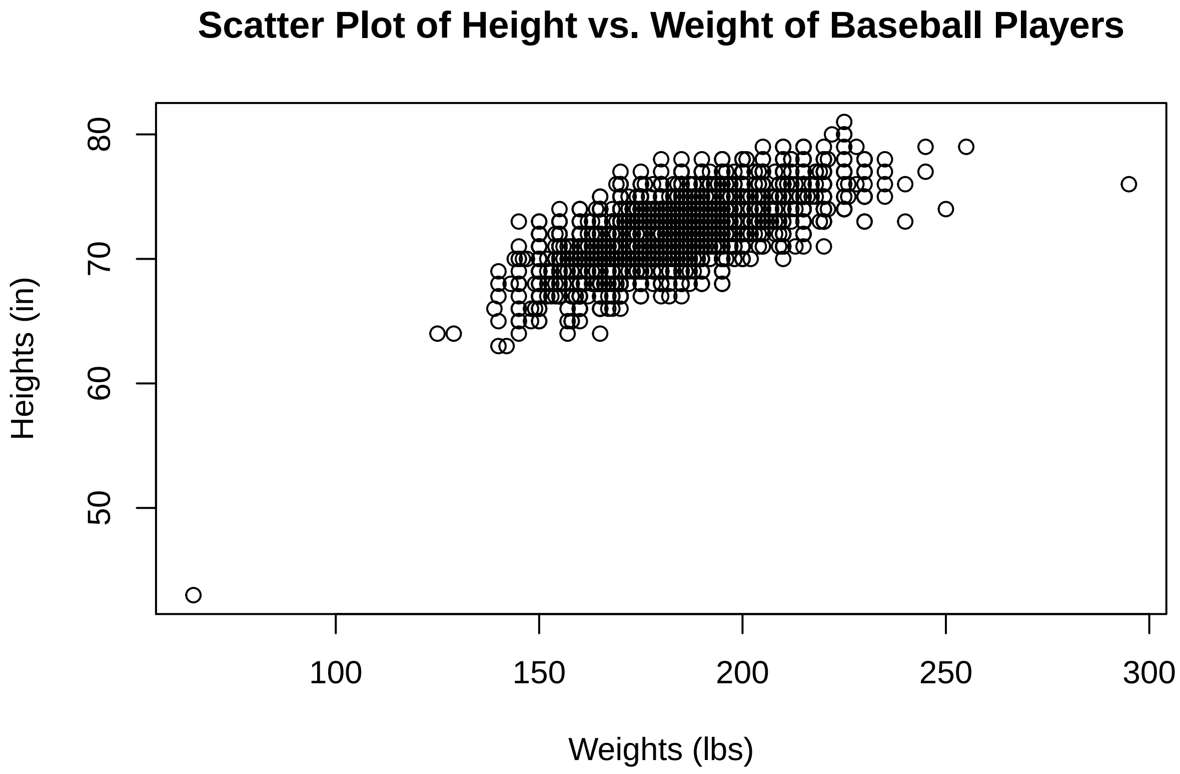


Answer: [Describe advantages of boxplots and histograms for this data and investigate usual features of the data]

The boxplot here tells us that the heights of baseball players, with the exception of one outlier, lie mostly close to the median height. The boxplot also allows us to see the summary statistics of the data, such as the quartiles, whereas the histogram allows us to see the distribution of the data (we can see that the heights are approximately normally distributed about a mean height of 72.23 inches, calculated above). The histogram also excludes the lowest height value, an outlier of 43 inches, as a result of its scaling, so the advantage of the boxplot is that the outlier is shown more clearly.

Part 2.3: (10 points) Create a scatter plot of the baseball player's heights as a function of their weight. Describe what the results show.

```r
weights <- players_born_1901_1950$weight
plot(weights, heights, main = "Scatter Plot of Height vs. Weight of Baseball Players",
     xlab = "Weights (lbs)", ylab = "Heights (in)")
```



Answer:

The height vs. weight scatterplot shows a positive correlation between height and weight. This makes sense since taller people tend to weigh more.

Problem 3: Examining categorical data

Let's now examine which states/regions baseball players are born in.

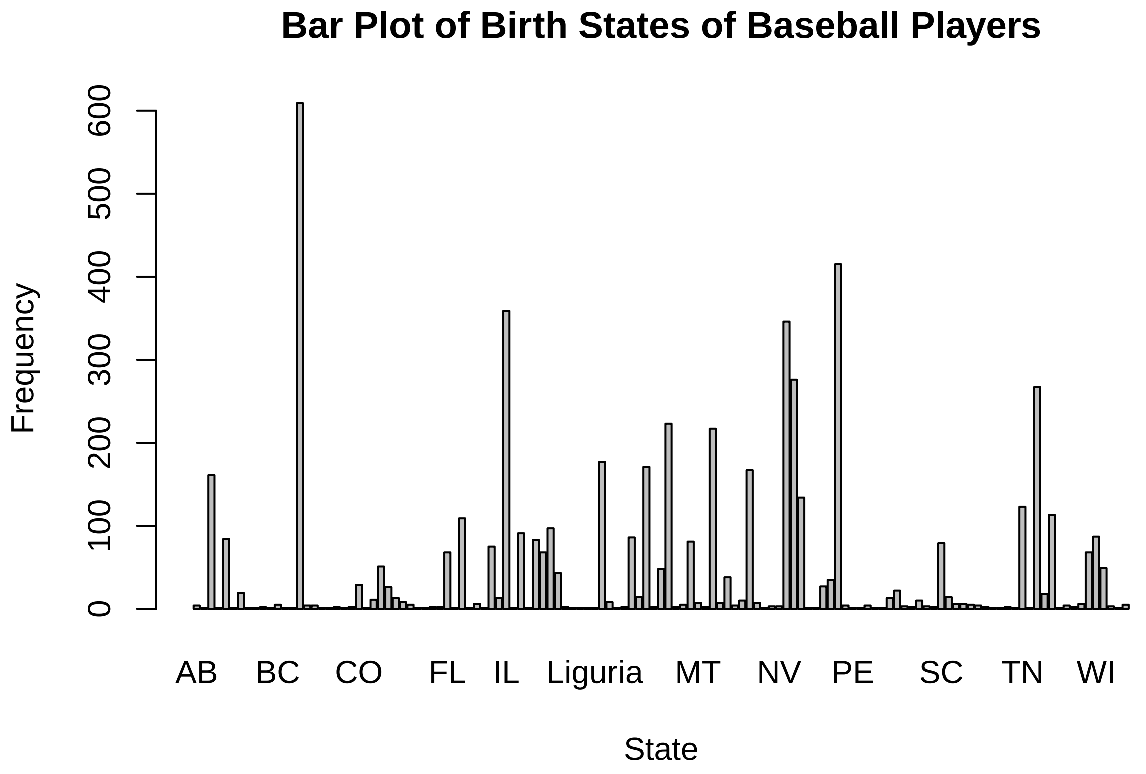
Part 3.1: (10 points) Use the table() function to create an object called birth_place_counts that has the counts of where players were born in. What is the state that the most players were born in?

Then create a bar plot and pie chart showing the counts of places that players are born in. How do these plots look? How could we make them better?

```
birth_states <- players_born_1901_1950$birthState
birth_place_counts <- table(birth_states)
sort(birth_place_counts, decreasing = TRUE)
```

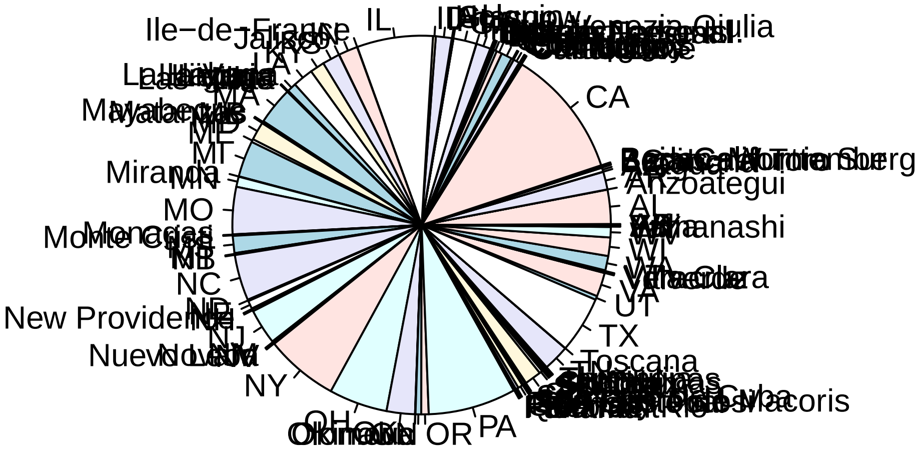
```
## birth_states
##                    CA                    PA                    IL
##                   609                   415                   359
##                    NY                    OH                    TX
##                   346                   276                   267
##                    MO                    NC                    MA
##                   223                   217                   177
##                    MI                    NJ                    AL
##                   171                   167                   161
##                    OK                    TN                    VA
##                   134                   123                   113
##                    GA                    LA                    IN
##                   109                    97                    91
##                    WI                    MD                    AR
##                    87                    86                    84
##                    KS                    MS                    SC
##                    83                    81                    79
##                    IA                    FL                    KY
##                    75                    68                    68
##                    WA                    CT                    WV
##                    68                    51                    49
##                    MN             La Habana                    NE
##                    48                    43                    38
##                    OR                    CO                    ON
##                    35                    29                    27
##                    DC                    RI                    AZ
##                    26                    22                    19
##                    UT                    ME                    SD
##                    18                    14                    14
##                    DE                    ID                    QC
##                    13                    13                    13
##                 Colon                    NH  San Pedro de Macoris
##                    11                    10                    10
##      Distrito Federal              Matanzas                    MT
##                     8                     8                     7
##                    ND                    NM                    HI
##                     7                     7                     6
##               Sinaloa                    SK                    VT
##                     6                     6                     6
##                    BC     Distrito Nacional          Monte Cristi
##                     5                     5                     5
##                Sonora                 Zulia                    AB
##                     5                     5                     4
##              Camaguey            Canal Zone        New Providence
##                     4                     4                     4
##                Panama         Pinar del Rio             St. Croix
##                     4                     4                     4
##              Veracruz            Nuevo Leon                    NV
##                     4                     3                     3
##         San Cristobal              Santiago                    WY
##                     3                     3                     3
##   Baja California Sur             Chihuahua            Cienfuegos
##                     2                     2                     2
##              El Seibo                Falcon               La Vega
##                     2                     2                     2
##                    MB               Miranda               Monagas
##                     2                     2                     2
##                    NB       San Luis Potosi      Santiago de Cuba
##                     2                     2                     2
##            St. Thomas            Tamaulipas           Villa Clara
##                     2                     2                     2
##                    AK            Anzoategui                Aragua
##                     1                     1                     1
##     Baden-Wurttemberg       Baja California              Barahona
##                     1                     1                     1
##                Berlin        Bocas del Toro              Carabobo
##                     1                     1                     1
##              Cheshire              Chiriqui              Coahuila
##                     1                     1                     1
##      Dodescanese Isl.                Duarte Friuli-Venezia Giulia
##                     1                     1                     1
##               Glasgow               Holguin         Ile-de-France
##                     1                     1                     1
##               Jalisco                  Lara            Las Villas
##                     1                     1                     1
##               Liepaja               Liguria             Mayabeque
##                     1                     1                     1
##                Novara               Okinawa               Olomouc
##                     1                     1                     1
##                    PE              Piedmont              Plzensky
##                     1                     1                     1
##                Puebla                 Sucre               Suffolk
##                     1                     1                     1
##             Thuringia               Toscana              Valverde
##                     1                     1                     1
##             Yamanashi
##                     1
```

```
barplot(birth_place_counts, main = "Bar Plot of Birth States of Baseball Players",
        xlab = "State", ylab = "Frequency")
```



```r
pie(birth_place_counts, main = "Pie Chart of Birth States of Baseball Players")
```

Pie Chart of Birth States of Baseball Players



Answers: The most players are born in California (we can see this when we sort the table, or after we create the bar and pie charts).

Both the bar plot and pie chart are oversaturated with data points, making the labels impossible to read (in the case of the pie chart) or missing (in the case of the bar plot, where there is not enough space for the labels). The plots should be reformatted to include only essential information (for example, taking only points with large enough values).

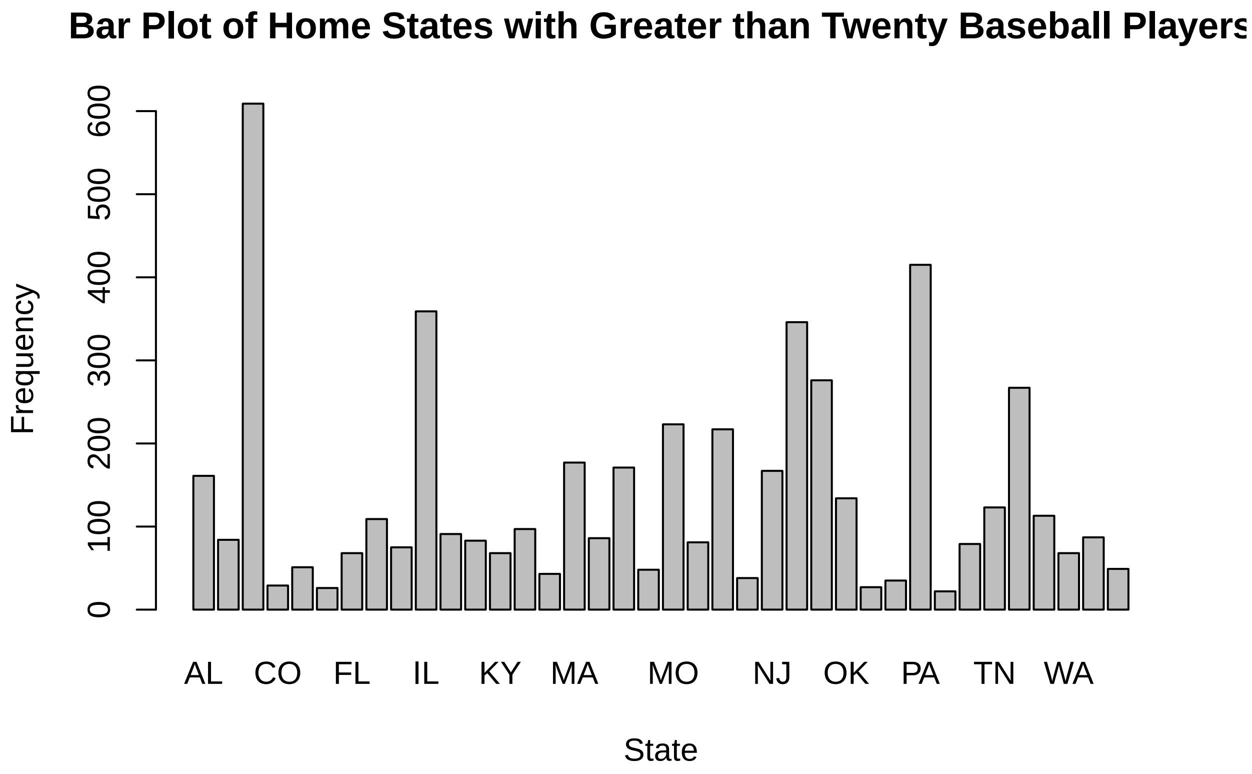
Part 3.2: (10 points)

Let's only plot states/places that have more than 20 players born in them. You can do this by creating a vector of booleans where TRUE indicates a state that has greater than 20 players born in it and FALSE indicates that 20 or less players were born in it (this can be done in 1 line of code). Then use this vector to extract only the places which more than 20 players born in. Finally replot the results with only states with more than 20 players born in them. Does this look better? Is there any place on this list that is not a state?

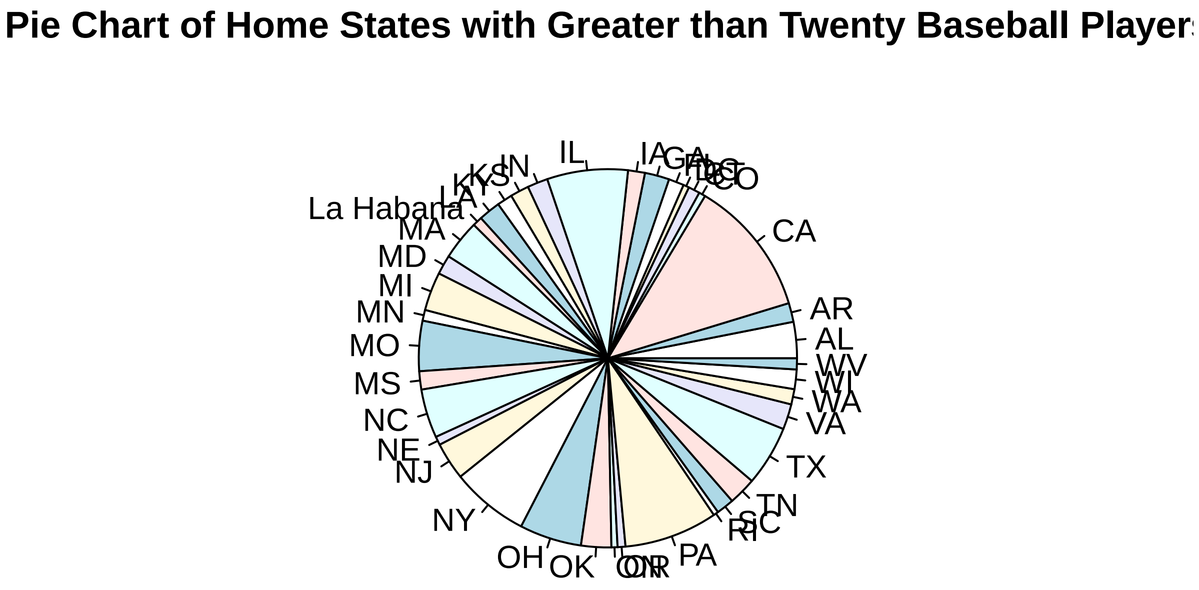
```
twenty_players_vec <- (birth_place_counts > 20)
twenty_players <- birth_place_counts[twenty_players_vec]
twenty_players
```

```
## birth_states
##         AL         AR         CA         CO         CT         DC         FL
##        161         84        609         29         51         26         68
##         GA         IA         IL         IN         KS         KY         LA
##        109         75        359         91         83         68         97
## La Habana        MA        MD        MI        MN        MO        MS
##        43       177        86       171        48       223        81
##        NC        NE        NJ        NY        OH        OK        ON
##       217        38       167       346       276       134        27
##        OR        PA        RI        SC        TN        TX        VA
##        35       415        22        79       123       267       113
##        WA        WI        WV
##        68        87        49
```

```
barplot(twenty_players,
        main = "Bar Plot of Home States with Greater than Twenty Baseball Players",
        xlab = "State", ylab = "Frequency")
```



```
pie(twenty_players,
    main = "Pie Chart of Home States with Greater than Twenty Baseball Players")
```



Answer: The data is clearer and easier to read, although some labels are still missing on the bar plot due to having too many bars, and labels on the pie chart are still difficult to read due to overlapping. The vector includes data for La Habana, a city in Cuba (not a state).

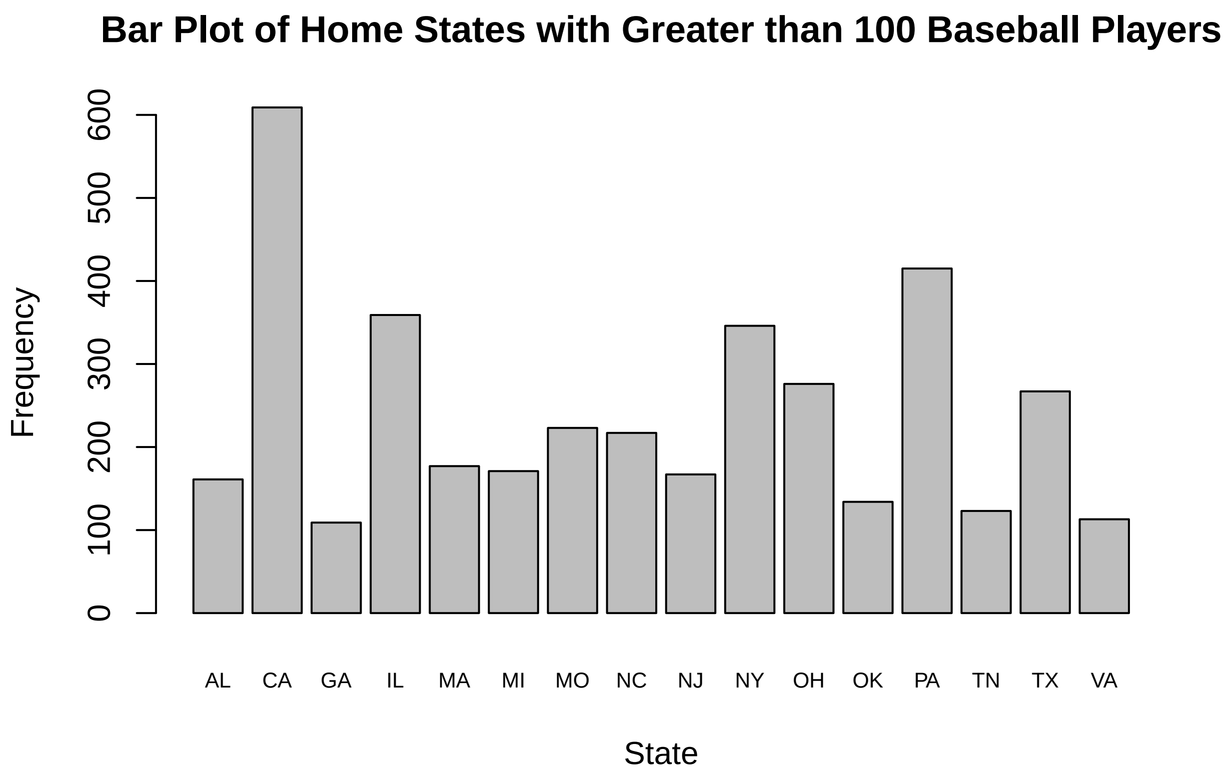
Part 3.3: (10 points)

The plots in part 3.2 still could look better. Adjust the plots so that you plot fewer states so that it is easier to see exactly which states most players are born in. Also adjust other visual attributes of the plots so that none of the labels are overlapping, and see if you can find other ways to make the plots look better, e.g., by adjusting the colors, etc. (hint: using ? pie and google will be helpful). Is plotting only some of the states misleading in any way, and if so, what are ways this could be addressed?

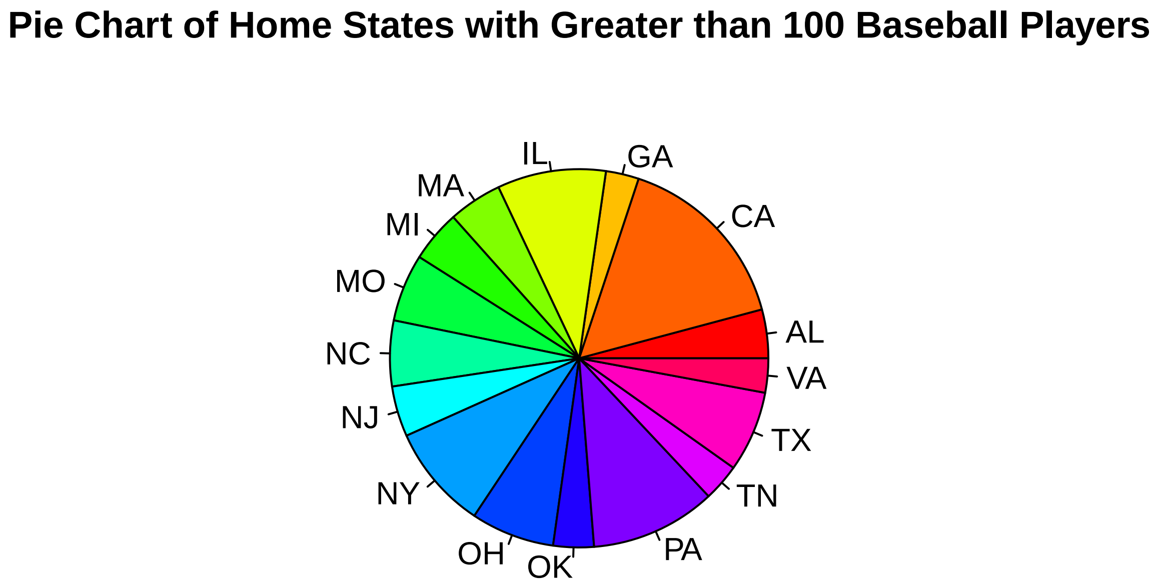
```
state_players_vec <- (birth_place_counts > 100)
state_players <- birth_place_counts[state_players_vec]
state_players
```

```
## birth_states
##  AL  CA  GA  IL  MA  MI  MO  NC  NJ  NY  OH  OK  PA  TN  TX  VA
## 161 609 109 359 177 171 223 217 167 346 276 134 415 123 267 113
```

```
barplot(state_players, main = "Bar Plot of Home States with Greater than 100 Baseball Players",
        xlab = "State", ylab = "Frequency", cex.names = 0.7)
```



```
pie(state_players, col = rainbow(length(state_players)),
    main = "Pie Chart of Home States with Greater than 100 Baseball Players")
```


Pie Chart of Home States with Greater than 100 Baseball Players

Answer:

The plot is misleading because it only includes 16 states of a total of 127 states (and cities in other countries), making it seem as if these states are the only birth states of baseball players, and the proportions of players from these states are significantly higher. To address this, we could create an "other" category including the total count from all of the excluded states, which would be one portion of the pie chart or one bar on the bar plot. This would make the percentages labeled in the pie chart accurate.

Problem 4: For loops (10 points)

As discussed in class, for loops allow you to repeat a process many times. Each time the process is repeated, a counter index object (usually named *i*) is incremented by 1. This is useful because it allows you to:

1. Repeat a process many times to generate results each time
2. Store each result in a vector using *i* to index into the vector.

The code below create uses a for loop to store the values of 1 squared up to 50 squared in a vector object named my_vec. Modify the code so that what is stored in the vector are the even integers from 2 to 100 (i.e, 2, 4, 6, . . . , 100).

```
my_vec <- NULL
for (i in 1:50){
  my_vec[i] <- i*2
}

my_vec
```

```
## [1]    2   4   6   8  10  12  14  16  18  20  22  24  26  28  30  32  34
## [18]  36  38  40  42  44  46  48  50  52  54  56  58  60  62  64  66  68
## [35]  70  72  74  76  78  80  82  84  86  88  90  92  94  96  98 100
```

Problem 5: Short reading (5 points)

As discussed in class, OkCupid is a dating website. One of the founders of the website, Christian Rudder, created a series of blog posts around 2010 where he analyzed data from the site to extract insights about dating. In order gain insight into what is possible from simple descriptive statistics and plots, please read the blog entry from July 7th 2010 title 'The Big Lies People Tell In Online Dating' and write one paragraph comment on something interesting you found in the article. Alternatively, you can read and comment on the article title 'How a Math Genius Hacked OkCupid to Find True Love' and comment on that article instead.

Describe something interesting you found in one of these articles:

After reading the first article published by OkCupid, I was unsurprised by many of their findings. The points published were all occurences I thought would be very common (adding height, exaggerating income, and using old pictures). One aspect I did find interesting was the method which the website used to draw their conclusions. For example, the blog post concluded that men on OkCupid are two inches shorter in reality than they list online by comparing the height distribution of U.S. men to the distribution of OkCupid users. While using the U.S. population to draw comparisons may be the most easily accessible method, I do not think these two data sets can be used to draw the conclusion they did (that the average male exaggerates by 2 inches). There is no indication that the average user exaggerates by a certain amount, and no comparison with their height in reality. I understand this data may be hard to collect, but without it we can only conclude that the OkCupid user is 2 inches taller than the average male.

Reflection (5 points)

Please reflect on how the homework went. In particular, please answer the following questions:

1. What concepts do you feel you are clearly understanding and which concepts are you confused about?
2. How many hours did you spend working on the homework?
3. How much did you enjoy doing the homework ("Super fun", "kind of fun", "not really", or "terrible")?
4. How much do you feel you learned doing this homework ("learned a lot", "learned some", "learned nothing", or "even more confused")?
5. Please note also if you went to TA office hours for help with this worksheet, and if the help you got was useful (in general, we strongly encourage you to attend TA office hours if you are having any difficulties with the homework).
6. Anything else you would like us to know?

Reflection Answers:

1. For me, the coding on this assignment is very clear and easy to understand. All of the concepts in the homework assignment were already covered in class. However, I am having trouble remembering statistics concepts and using statistical terms to describe the data, even though the knowledge required for this assignment was still very basic.
2. About 2.5 hours
3. Kind of fun!
4. Learned some - it was good to review and apply the concepts and code we went over in class, otherwise I wouldn't have been able to retain the information.
5. I got help from Duda, which helped me address a few details I missed (such as labeling or sorting tables).
6. Thank you Prof. Meyers - I'm enjoying the class so far and feel that it is moving at a reasonable pace. The lectures are clear and the lecture slides are even more helpful.